Introduction to the Minitrack on Data Analytics, Data Mining, and Machine Learning for Social Media

Jeffry S. Babb
West Texas A&M University
jbabb@wtamu.edu

Kevin D. Mentzer
Bryant University
kmentzer@bryant.edu

David J. Yates
Bentley University
dyates@bentley.edu

Abstract

Social media continues to reflect our daily lives. Ubiquitous mobile devices allow for immersion into social media on a nearly universal basis. As such, the blurring of the line between our online and offline worlds and personas only continues to increase. Traditional text-based platforms are increasingly turning to image and video content, with video-intensive platforms appealing more to younger generations. We are also starting to see longer threads, stories, and conversations emerging, meaning there is significance in looking at social media messages as part of a larger discussion. These shifts result in both additional challenges and opportunities for researchers to understand the role social media is playing in society. The papers in our minitrack represent the diversity of our field. We see data gathered from multiple platforms as well as new techniques to analyze these data.

1. Sessions and papers at a glance

This minitrack contributes nine papers to the fifty-fifth HICSS. These papers are described in groups of three in the following three paragraphs.

In Shahi, Clausen, and Steiglitz's "Who shapes crisis communication on Twitter? An analysis of German influencers during the COVID-19 pandemic," we are entreated with empirical investigations of globally-relevant issues related to public health and responses to a crisis. These authors explore how Twitter reflects and shapes the public response to the first wave of the COVID-19 pandemic in Germany. They show that news and journalist accounts were influential throughout this first wave. However, government accounts were particularly important before, during, and after high-impact governmental responses, e.g., the lockdown in March and April of 2020. Okpala et al., in their paper "Perception Analysis: Pro- and Anti-Vaccine Classification with NLP and Machine Learning," remind us how regional analysis of social media can be a viable way to assess public perception and to identify actionable data in a targeted context. This study explores how to use Natural Language Processing (NLP) and machine learning to understand perceptions of COVID-19 vaccination in the Greater Cincinnati area as the first vaccines became available. An overview of the authors' model for data categorization and machine learning is shown in Figure 1.

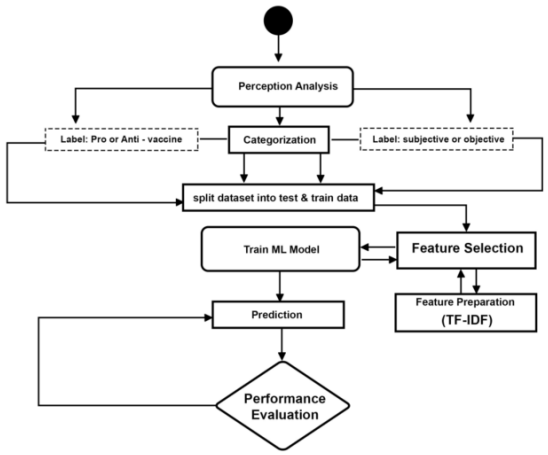

Figure 1. Vaccine Data Categorization and Machine Learning Model (Source [1]).

We also have empirical studies examining how to better understand and navigate individual symptoms of chronic disease. Van Hoven, Thoms, and Botts investigate this area with their paper "Determining Link Relevancy in Tweets Related to Multiple Myeloma Using Natural Language Processing," which uses an innovative combination of web mining, Twitter hashtags, and NLP to discover and aggregate information about the symptoms of Multiple Myeloma. The authors propose and evaluate an algorithm for polling the Twitter API for relevant tweets, extracting relevant metadata, and assessing the relevancy of these tweets based on their content.

Schaal, Davis, and Mueller adopt a holistic perspective concerning the impacts of social media as it pertains to the policy-making process in their paper "Multi-National Topics Maps for Parliamentary Debate Analysis." Using Germany, Spain, and the United Kingdom as examples, this paper explores how to join the corpora of political debate in these countries into a

URI: https://hdl.handle.net/10125/79694
978-0-9981331-5-7

single corpus and use probabilistic topic modeling on this corpus to create a reference topic model for topic linking. The authors show the dynamic nature of topical discourse over time in different parliaments, distinguishing between topics shared between parliaments, specific to one country, or neither. Koss and Bohnet-Joschko investigate what can be learned about patients' unmet needs for Multiple Sclerosis (MS) treatment using social media data in their paper "Social Media Mining in Drug Development Decision Making: Prioritizing Multiple Sclerosis Patients' Unmet Medical Needs." This study adapts the Opportunity Algorithm (OA, from outcome-driven innovation) to identify and prioritize unmet medical needs of MS patients shared in social media posts. Features for the OA are generated using topic modeling and sentiment analysis. The results demonstrate the potential of this method to provide relevant insights into rare disease populations to improve patient-centered drug development. Other insights are offered in explorations concerning individuals and their deepening awareness of the world around them. In an appraisal of the impact of continued innovation and product development in the smart speaker market, Shim, Lo, and Liew share their text-mining sentiment analysis of consumer reviews to understand expressions of consumer preference on social media. Their paper is entitled "Do Sequels Outperform or Disappoint? Insights from an Analysis of Amazon Echo Consumer Reviews." Shim et al.'s approach identifies which product features outperform or disappoint, and whether positive or negative sentiment increases or decreases over time. Insights from such analysis are especially important as the new generation of a product is announced and released.

Questions related to daily life are also the subject of the remaining papers in this minitrack. Social media is often used as a place to seek advice and calibrate perspectives related to prevailing sentiments. Sometimes the advice sought is of a personal nature, and speaks to our relationships. Cannon et al. explore this space in communities on Reddit in their paper "'Don't Downvote A$$$$$$s!!': An Exploration of Reddit's Advice Communities." These authors analyze a dataset with top posts from two Reddit forums that discuss relationships, r/AmItheAsshole and r/relationships, and extract natural language features, including sentiment, similarity, word frequency, and demographics. Their results point to the need for research that includes a more careful, inclusive focus on identity and disclosure and how these interact with advice-seeking behavior in online communities. As a multifaceted resource, social media can be a source of as much disinformation as information, increasing the complexities that accompany the benefits of our interconnectedness. Koohikamali and Gerhart examine the phenomena related to a specific event to model flows of information, rumor, and disinformation in their paper "False Rumor (Fake) and Truth News Spread During a Social Crisis." This paper tests a research model to examine significant determinants of message spread during the 2016 Charlotte, North Carolina protests which occurred after false online rumors spread related to the shooting of Keith Lamont Scott (see Figure 2).

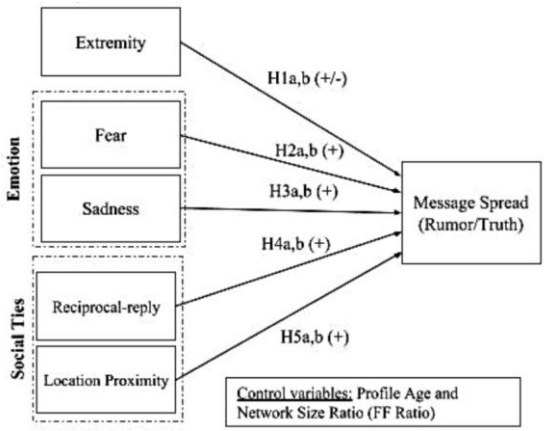

Figure 2. Proposed Rumor/Truth Diffusion Model During a Social Crisis (Source [2]).

The results of this study provide theoretical and practical insights into the current research in information diffusion and social engagement. Finally, at a fundamental level, all of our digitally-mediated activities leave footprints that are discernable and traceable, but are challenging to retrace and search. An approach to comprehending these footprints is explored in our last paper, "A Frequency-Based Learning-To-Rank Approach for Personal Digital Traces." In this research, Vianna and Marian propose and evaluate a multidimensional data model for personal data based on six natural questions — what, when, where, who, why, and how — and a novel learning-to-rank approach that maps these questions to features. Their evaluation shows that a frequency-based learning approach improves search accuracy when compared with traditional search tools.

References

[1] I. Okpala, G. Romera Rodriguez, W. Zheng, S. Halse, and J. Kropczynski, "*Perception Analysis: Pro- and Anti-Vaccine Classification with NLP and Machine Learning*," in Proc. of 55th Hawaii International Conference on System Sciences, Lahaina, HI, USA, January 2022.

[2] M. Koohikamali, and N. Gerhart, "*False Rumor (Fake) and Truth News Spread During a Social Crisis*," in Proc. of 55th Hawaii International Conference on System Sciences, Lahaina, HI, USA, January 2022.